ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

A hybrid feature selection and ensemble learning model for heart disease diagnosis

Anurag Tripathi[1], Rahul Vyas[1], Gaurav Dwivedi[2]*, Lalit Kumar Tripathi[3], Ajai Kumar Maurya[3], and Ali Zaheer Agha[3]

[1]Department of Computer Application, United University, Prayagraj, Uttar Pradesh, India

[2]Department of Computer Science and Engineering, United University, Prayagraj, Uttar Pradesh, India

[3]Department of Computer Science and Engineering, United College of Engineering and Research, Prayagraj, Uttar Pradesh, India

Abstract

Heart disease remains one of the leading causes of mortality worldwide, making early and accurate diagnosis essential for improving patient outcomes. This study proposes a hybrid feature selection and ensemble learning framework for heart disease diagnosis using a publicly available dataset containing 303 records and 13 clinical attributes. The preprocessing pipeline included missing-value imputation, one-hot encoding of categorical variables, and minimum–maximum normalization. Support vector machine–recursive feature elimination (SVM–RFE) was used to identify the most informative clinical features and remove redundant attributes. Multiple machine learning classifiers, including logistic regression, k-nearest neighbors, SVMs, decision trees, random forests, and extreme gradient boosting (XGBoost), were evaluated. Bayesian optimization was applied to tune model hyperparameters, and performance was assessed using an 80:20 train–test split together with stratified five-fold cross-validation. The proposed ensemble model achieved 93.62% accuracy, 94.08% precision, 93.11% recall, 93.58% F1-score, and 0.96 receiver operating characteristic area under the curve. The selected features included chest pain type, ST-slope, exercise-induced angina, maximum heart rate, and age, all of which are clinically meaningful predictors of cardiovascular risk. The results show that combining SVM-RFE, hyperparameter optimization, and ensemble learning improved predictive performance, generalization, and interpretability. The proposed framework can support early heart disease screening and clinical decision-making.

Keywords: Heart disease diagnosis; Electronic health records; Feature selection; Ensemble learning; Bayesian optimization; Machine learning

***Corresponding author:**
Gaurav Dwivedi
(gaurav.dwivedi@uniteduniversity.edu.in)

Citation: Tripathi A, Vyas R, Dwicedi G, Tripathi LK, Maurya AK, Agha AZ. A hybrid feature selection and ensemble learning model for heart disease diagnosis. *Artif Intell Health*. 2026;3(4):026200045. doi: 10.36922/AIH026200045

Received: May 15, 2026

Revised: June 24, 2026

Accepted: June 25, 2026

Published online: July 14, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Recent advances in machine learning and artificial intelligence (AI) have significantly improved the early diagnosis of cardiovascular diseases using electronic health record (EHR) data. Researchers have explored various classification algorithms, feature selection techniques, and ensemble learning approaches to enhance predictive performance. This

section reviews representative studies related to heart disease prediction, feature selection, hyperparameter optimization, and ensemble learning, providing the foundation for the proposed framework.[1] Chronic diseases represent a major global public health challenge and are responsible for a substantial proportion of worldwide morbidity and mortality. One of the greatest challenges facing world health systems is the need to manage chronic diseases, and their impact on nations' health is immense.[2] These conditions contribute significantly to healthcare expenditure, reduced quality of life, and increased mortality worldwide. These include disorders such as diabetes, renal disease, cancer, and cardiovascular disease. The demand to develop innovative and robust predictive models to identify early in the context of individualized therapies is urgent, as the number of chronic diseases is increasing steadily due to lifestyle changes and an aging population. Identifying chronic diseases at their initial stages is instrumental in preventing complications, reducing disease severity, and preventing further deterioration among affected persons.[3] Much of the nature of various chronic illnesses, such as heart disease, diabetes, and certain malignancies, is that they exhibit no symptoms in their initial stages; thus, slow detection and acceptance by administrators are postponed until bitter, potentially irreversible, steps are taken. AI has become a significant domain of progress recently and has affected medical and healthcare science in phenomenal ways.[4] AI can identify patients at high risk of developing certain diseases, enabling the introduction of early intervention and prevention plans. Such capacity for transformation allows healthcare professionals to detect chronic diseases at early stages, thus reducing the burden on patients and healthcare systems through ongoing education.[5] AI models can adapt to new data and provide individualized, accurate predictions that help medical practitioners develop a tailored treatment strategy based on the risks affecting a particular individual. AI techniques enable the extraction of meaningful patterns and predictive insights from large-scale healthcare data. Machine learning algorithms have demonstrated significant potential for identifying hidden relationships and predictive patterns within complex medical datasets. Latent structures in complicated medical data.[6]

These capabilities can improve diagnostic accuracy, support treatment planning, and enhance overall patient outcomes. As the volume of medical data consumed has been exponentially increasing, researchers have explored a range of machine learning techniques to improve early detection and prediction. All this effort is part of a global effort to fight chronic disease at a broader level. Extreme gradient boosting (XGBoost) is a prominent machine learning algorithm that is based on the gradient-boosting decision tree algorithm.[7] XGBoost has demonstrated strong predictive performance, excellent generalization capability, and robustness across a wide range of healthcare prediction tasks. It has also demonstrated outstanding performance across several industries, including healthcare. Although the XGBoost algorithm has significant strengths, it may perform poorly when parameter tuning is inadequate in its default settings. Suboptimal hyperparameter settings may reduce predictive performance and limit the model's ability to generalize to unseen data, thereby impairing its generalization and adaptation capacities. In the present research, Bayesian optimization was used to successfully tune hyperparameters and improve model performance. The use of Bayesian optimization in hyperparameter optimization is central to achieving the best performance, avoiding overfitting, and improving the model's robustness. XGBoost and Bayesian optimization are two promising hybrid approaches that, when combined, yield a comprehensive and accurate chronic illness predictor.[8]

Given the difficulty of handling large medical datasets with many features, we propose a new method for heart disease diagnosis using EHR data, starting with feature engineering and using support vector machines (SVMs) and recursive feature elimination (RFE). The iterative nature of RFE progressively eliminates less impactful variables based on SVM weights, yielding a subset of the most relevant features. Subsequently, the refined dataset is fed into the powerful hybrid Bayesian optimization–XGBoost model to predict chronic diseases, accounting for the impact of symptoms on disease presence.[9] The performance of the proposed framework is evaluated using multiple classification metrics, including accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC–AUC).[10]

1.1. Research contributions

The major contributions of this study are summarized as follows:

(i) A hybrid feature-selection framework based on SVM–RFE is developed to identify the most relevant clinical attributes associated with heart disease.
(ii) An ensemble learning model is proposed to improve diagnostic performance by combining the predictive strengths of multiple machine learning classifiers.
(iii) Bayesian optimization is employed to optimize model hyperparameters and enhance predictive robustness and generalization capability.
(iv) A comprehensive experimental evaluation is conducted using accuracy, precision, recall, F1-score, and ROC–AUC to assess the effectiveness of the

proposed framework.

(v) Explainability and feature-importance analysis are incorporated to improve the interpretability of model predictions and support clinical decision-making.

(vi) The proposed framework provides an efficient and reliable decision-support approach for early heart disease diagnosis using EHR data.

1.2. Research gap

Recent studies have shown that feature selection, ensemble learning, and hyperparameter optimization can improve heart disease prediction. However, many existing works rely on a single validation strategy, report limited interpretability, or do not clearly explain how redundant features are removed before model training. Some studies also present high accuracy without sufficiently addressing generalizability or overfitting. To address these limitations, the present study integrates SVM–RFE, Bayesian optimization, and ensemble learning in a unified framework for heart disease diagnosis, with explicit validation and interpretability analysis.

2. Related works on heart disease prediction and machine learning-based diagnosis

Although several studies have investigated machine learning applications across various chronic diseases, including chronic kidney disease (CKD), diabetes, and cancer, this section focuses on methodological advances that are transferable to heart disease diagnosis, particularly feature selection, ensemble learning, and hyperparameter optimization.

A convolutional neural network (CNN)–multicast-based dissemination and reroute planning (MDRP) approach, which can combine structured and unstructured hospital data in a massive aggregation, was proposed by Wu *et al.*[11] to extract features from unstructured data. The Naïve Bayes technique was trained and evaluated. Better prediction capabilities, with an accuracy of 94.2%, were the results. The primary limitation of their study is that the model is inadequate for intricate conditions such as neurological disorders. Jongbo *et al.*[12] developed a clinical data-based decision-making model to predict diseases using historical information, considering non-training patterns and incorporating them into the model. Their study also used data on various disorders, achieving an accuracy of 94.5%. The primary limitation of their study is its failure to account for feature reduction. Hannan and Pal[13] introduced a machine learning method for predicting chronic diseases, using real-time data from Chinese hospitals and reconstructing missing data using a latent factor model. The proposed framework utilized multimodal healthcare data and employed CNNs for predictive modeling. The chosen ailment was a brain fracture. The study reported an accuracy of 94.8%, demonstrating the effectiveness of multimodal data integration for disease prediction. Chicco *et al.*[14] conducted a comprehensive investigation on a liver dataset using several machine learning techniques, including fuzzy logic, fuzzy neural networks, and decision trees. They found that using a fuzzy neural network to classify data had a 91% success rate.

A classification method that builds upon standard regression is logistic regression. This approach is limited to representing binary variables, such as whether an event occurs. It is useful for determining the likelihood that a newly created instance belongs to a certain class cluster.[15] The prediction of chronic illnesses has become increasingly significant within the research community, creating several opportunities for study and development. Naïve Bayes is a classification method used by the research community, primarily for training and classification, and is fundamentally based on Bayes' theorem.[16] Using the information gleaned from the training model relevant to the circumstances of an event, this theorem calculates the likelihood of the event occurring. The basic idea behind this classifier is that characteristics within a class are interdependent among themselves[17] if a specific class characteristic is unconnected to any other feature of the same class. Decision trees are effective for both decision-making and predictive analytics. By applying decision logic, a decision tree model can evaluate data and classify it according to predetermined criteria, much like a tree. There is a hierarchical structure with the root node at the highest level and each node below it.[18] Kong *et al.*[19] proposed a latent component model for the reconstruction of health care data on cerebral infarction. The model can handle incomplete data, which may occur in healthcare.

Basha *et al.*[20] proposed an ensemble learning solution for CKD diagnosis by combining various classifiers to improve prediction accuracy. They showed that an ensemble of machine learning classifiers improves diagnostic performance as opposed to single models. Their study emphasized the usefulness of ensemble techniques for diagnosing medical conditions and, more importantly, for managing complex patient data.

Sharmila *et al.*[21] aimed to optimize SVMs using information gain and AdaBoost to improve CKD diagnosis. Their strategy was to increase model precision by selecting the most valuable features from patient data sets. The experiment revealed that a hybrid technique achieved better results than a standard SVM, demonstrating the importance of feature selection and boosting methods in medical prediction.

Ruiz *et al.*[22] examined the application of CNNs to identify and classify CKD, a major concern in medical diagnostics, because they can process structural data and identify minor patterns in medical data.

Liu *et al.*[23] analyzed EHRs of CKD subjects using machine learning algorithms to identify who was at the highest risk of cardiovascular disease (CVD). They demonstrate that CKD is a risk factor for cardiovascular disease and illustrate how predictive models can support its early management. In their study, several categorization algorithms were employed, and the results were compared using industry-standard metrics, including recall, accuracy, and precision. They proposed a data-informed method for early CKD diagnosis that leverages an explainable AI (XAI) data-copying model. The study focused on the ethics of AI decision-making, with healthcare professionals able to understand the predictions. The model incorporates state-of-the-art feature selection and machine learning practices to be more interpretable and accurate, and to apply better in the real clinical environment.

3. Materials and methods

3.1. Study design

This study presents a machine learning-based framework for the early diagnosis of heart disease using EHR data. The proposed framework integrates feature selection, hyperparameter optimization, and ensemble learning techniques to improve predictive performance and diagnostic reliability. A systematic experimental pipeline was developed to preprocess patient data, identify the most relevant clinical features, optimize model parameters, and evaluate multiple machine learning classifiers for heart disease prediction.

The overall workflow of the proposed framework consists of five major stages: (i) data acquisition, (ii) data preprocessing, (iii) feature selection, (iv) model development and optimization, and (v) performance evaluation. The development and reporting of the proposed framework were aligned with Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD)–AI principles to improve transparency, reproducibility, and methodological rigor.

3.2. Dataset description

The study used a publicly available heart disease dataset containing patient demographic and clinical information associated with cardiovascular conditions. The dataset comprises multiple attributes commonly used in clinical diagnosis, including age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression, slope of peak exercise ST segment, number of major vessels, and thalassemia status.[24]

The target variable indicates whether a patient has heart disease. Each patient record was assigned to one of the two diagnostic classes, enabling the development of a binary classification framework for heart disease prediction. Table 1 presents the clinical attributes utilized for heart disease prediction.

Table 1. Description of clinical attributes used for heart disease prediction

Feature	Description
Age	Age of patient in years
Sex	Gender of patient (male/female)
Chest pain type (cp)	Type of chest pain experienced
Resting blood pressure (trestbps)	Resting blood pressure measured in mmHg
Cholesterol (chol)	Serum cholesterol level (mg/dL)
Fasting blood sugar (fbs)	Blood sugar > 120 mg/dL
Resting ECG (restecg)	ECG results
Maximum heart rate (thalach)	Maximum heart rate achieved
Exercise-induced angina (exang)	Exercise-induced chest pain
ST depression (oldpeak)	ST depression induced by exercise
Slope	Slope of peak exercise ST segment
Coronary arteries (CA)	Number of major vessels colored by fluoroscopy
Thalassemia (thal)	Thalassemia status
Target	Presence or absence of heart disease

Abbreviation: ECG: Electrocardiography.

3.3. Inclusion and exclusion criteria

To ensure data quality and reliability, inclusion and exclusion criteria were applied before model development:

(i) Inclusion criteria:
- Patient records containing complete diagnostic labels.
- Clinical attributes relevant to cardiovascular disease diagnosis.
- Records with sufficient demographic and physiological information.

(ii) Exclusion criteria:
- Duplicate patient records.
- Records containing inconsistent or invalid values.
- Features with excessive missing information.
- Non-clinical identifiers that do not contribute to disease prediction.

The application of these criteria ensured the generation of a high-quality dataset suitable for machine learning analysis (Table 2).

3.4. Data preprocessing

Data preprocessing was performed to improve data quality and ensure compatibility with machine learning algorithms. The preprocessing pipeline consisted of the following steps:

(i) Missing value treatment: Missing numerical values were replaced using mean imputation, while categorical attributes were imputed using mode values. This strategy minimizes information loss while preserving the statistical properties of the dataset.
(ii) Categorical data encoding: Categorical variables were transformed into numerical representations using one-hot encoding to facilitate algorithmic processing.
(iii) Data normalization: Feature values were normalized using minimum–maximum (Min–Max) scaling to ensure that all variables contributed equally during model training.

The normalized feature value was calculated as:

$$X_{norm} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad (1)$$

where X represents the original feature value, X_{min} denotes the minimum value, and X_{max} denotes the maximum value of the feature.

3.5. Feature selection using support vector machine–recursive feature elimination

High-dimensional datasets often contain redundant and irrelevant features that negatively affect classification performance. Therefore, SVM–RFE was employed to identify the most informative clinical attributes.[25] SVM–RFE is a wrapper-based feature selection approach that iteratively removes features with the lowest contribution to classification performance. At each iteration, the SVM model assigns weights to all features, and the least important feature is eliminated. The process continues until the optimal subset of features is obtained. The selected features were subsequently used for model training and evaluation. The primary advantages of SVM–RFE include:

- Reduced feature redundancy.
- Improved classification accuracy.
- Enhanced computational efficiency.
- Improved model interpretability.

Table 2. Inclusion and exclusion criteria applied to the dataset

Category	Criteria
Inclusion criteria	Patient records containing complete heart disease diagnosis labels
	Clinical features relevant to cardiovascular disease prediction
	Records containing demographic and physiological attributes
	Records suitable for machine learning analysis after preprocessing
Exclusion criteria	Duplicate patient records
	Records containing inconsistent or erroneous values
	Features with excessive missing values
	Non-clinical identifiers such as patient IDs
	Incomplete records that could not be reliably imputed

3.6. Ensemble learning framework

The proposed diagnostic framework employs ensemble learning to improve prediction robustness and generalization capability.[26] Multiple machine learning classifiers were considered during model development, including:

- Decision tree
- Logistic regression
- Random forest
- XGBoost

The ensemble framework combines the predictive capabilities of individual classifiers to generate more reliable diagnostic outcomes. By integrating multiple learning algorithms, the framework reduces model variance, minimizes overfitting, and improves prediction stability.

3.7. Hyperparameter optimization

The predictive performance of machine learning algorithms is strongly influenced by hyperparameter configuration. Therefore, Bayesian optimization was employed to identify optimal parameter settings for each classifier.[27] The objective was to maximize predictive performance while minimizing model complexity and overfitting. Bayesian optimization iteratively explores the search space and selects parameter combinations that maximize model performance while minimizing computational cost.[28]

The optimized parameters included:

- Number of estimators
- Learning rate
- Maximum tree depth
- Minimum samples per split
- Regularization parameters

The optimization process ensured improved model generalization and reduced the likelihood of overfitting (Table 3).

3.8. Validation strategy

To ensure reliable performance estimation and prevent overfitting, a stratified k-fold cross-validation strategy was employed. The dataset was initially split into training and test sets at an 80:20 ratio. Subsequently, k-fold cross-validation was performed on the training set to evaluate model stability and robustness.[29] Stratified sampling was used to preserve class distribution across all folds. This validation strategy provides a more reliable estimate of model performance and minimizes the risk of data leakage. The development, validation, and reporting of the proposed machine learning framework were guided by the TRIPOD-AI reporting recommendations to enhance transparency and reproducibility.

3.9. Performance evaluation metrics

The effectiveness of the proposed framework was assessed using widely accepted classification metrics:

(i) Accuracy: Accuracy measures the proportion of correctly classified instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

(ii) Precision: Precision measures the proportion of correct positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

(iii) Recall: Recall measures the proportion of actual positive instances correctly identified.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

(iv) F1-score: The F1-score represents the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{IV}$$

(v) ROC–AUC: ROC–AUC evaluates the model's ability to distinguish between positive and negative classes across different classification thresholds.[30] A higher ROC–AUC value indicates superior discriminatory capability and improved diagnostic performance (Table 4).

Although previous studies have reported promising results for heart disease prediction, several limitations remain. Many existing approaches rely on individual classifiers, limited feature-selection strategies, or insufficient validation procedures. In addition, several studies focus primarily on predictive accuracy, neglecting model interpretability, robustness, and generalization. Furthermore, integrating SVM–RFE-based feature selection, Bayesian hyperparameter optimization, and ensemble learning within a unified framework remains relatively unexplored for heart disease diagnosis. Therefore, this study proposes a hybrid framework that combines these components to improve predictive performance while maintaining interpretability and clinical relevance.

Table 3. Hyperparameter settings used for model optimization

Model	Hyperparameter	Search range/Value
Decision tree	Max depth	3–20
Decision tree	Min samples split	2–10
Decision tree	Criterion	Gini, Entropy
Random forest	Number of trees (n_ estimators)	100–500
Random forest	MaxdDepth	5–30
Random forest	Min samples split	2–10
Logistic regression	Regularization parameter ©	0.001–100
Logistic regression	Solver	lbfgs, liblinear
XGBoost	Number of trees (n_ estimators)	100–500
XGBoost	Learning rate	0.01–0.30
XGBoost	Max depth	3–15
XGBoost	Subsample	0.6–1.0
XGBoost	Colsample_bytree	0.6–1.0

Abbreviation: XGBoost: Extreme gradient boosting.

3.10. Major heart disease risk factors

Heart disease remains one of the leading causes of mortality worldwide and represents a major public health challenge. Cardiovascular diseases encompass a broad range of disorders affecting the heart and blood vessels, including coronary artery disease, heart failure, arrhythmias, and myocardial infarction. Early identification of high-risk patients is critical for reducing mortality, preventing complications, and improving clinical outcomes. Machine learning techniques have recently emerged as promising tools for supporting early diagnosis by analyzing EHR data and other clinical indicators.[31] Major heart disease risk factors are presented in Table 5.

4. Learning ensemble framework

Ensemble learning has been widely adopted across diverse application domains, including healthcare, finance, fraud detection, image classification, and natural language processing. Ensemble learning is especially efficient in diagnosing chronic diseases because it can process high-dimensional EHRs, discover informative patterns, and make more accurate and trustworthy predictions.[32] The framework has several important elements, including base learners (weak models), ensemble strategies, and aggregation mechanisms. Base learners are individual machine learning models that serve as the foundational components of an ensemble framework, including, e.g., decision trees, SVMs, neural networks, logistic regression, random forests, XGBoost, and Naïve Bayes. To improve overall performance, such low-level learners may be trained independently on different data collections or learning paradigms, and their predictions combined. Ensemble strategies combine base models, and among the most widely used inductive ensemble methods are bagging, boosting, stacking, and voting. Bagging bagging/slang can be used to solve regression tasks. It (also referred to as bootstrap aggregating) involves training multiple models on randomly selected subsets of data.[33]

With classification, bagging involves averaging multiple predictions from each model. Random forests are a common bagging method in which a large number of decision trees are trained on different subsets of the dataset, and the individual predictions of the trees are later averaged. In contrast, training models sequentially, where each model is trained on the errors of the preceding model, results in a more accurate solution. These algorithms include, but are not limited to, AdaBoost and gradient boosting, XGBoost, and light gradient boosting machine (LightGBM). In a medical appliance, boosting is an excellent strategy, for example, for the prediction of heart disease, diabetes, or CKD, because it reduces the number of unsuccessful classifications by correcting the weights and concentrating on wrong classifications. Stacking (stacked generalization) is a type of ensemble in which many base models are trained, and their predictions are used as inputs to a meta-learner to produce a final result. Unlike bagging and boosting, which use homogeneous models, stacking can combine heterogeneous models, such as neural networks, XGBoost, and SVMs, to maximize predictive performance. One of the areas where stacking has been utilised well is in chronic disease diagnosis. In this method, a large number of models are evaluated based on measurements of various elements of a patient's health information, and the final layer integrates their outputs to provide a trustworthy diagnosis.[34] When voting on classification problems, the ensemble method employs a plurality vote; for regression problems, it uses plurality averaging. Voting is of two kinds: hard voting, in which the final prediction depends on the majority class label in each model, and soft voting, in which

Table 4. Summary of key studies on machine learning and optimization for disease prediction

Author(s) & year	Study focus	Methodology/Algorithm used	Key contribution
Jain & Agrawal (2023)[18]	Heart failure prediction	XGBoost, LR, support vector classifier	Compared multiple ML models for heart disease prediction
Kong *et al.* (2023)[19]	Credit scoring optimization	XGBoost with SHAP and Bayesian optimization	Improved model interpretability for financial risk assessment
Wu *et al.* (2019)[11]	Hyperparameter tuning	Bayesian optimization for ML models	Enhanced efficiency in hyperparameter selection
Bergstra & Bengio (2012)[10]	Hyperparameter optimization	Random search method	Showed random search can outperform grid search
Al-Jamimi (2022)[3]	AI in petroleum industry	Neural networks	Modeled catalytic desulfurization using AI
Basha *et al.* (2023)[20]	Industrial process optimization	Bayesian-optimized neural networks	Applied AI to gas-to-liquid plant modeling
Sharmila *et al.* (2017)[21]	Disease classification	Comparative study of ML algorithms	Evaluated different classifiers for disease prediction
Burse *et al.* (2019)[4]	Heart disease prediction	Neural networks with preprocessing	Studied impact of data preprocessing on neural network performance
Hosmer *et al.* (2013)[9]	Statistical modeling	LR	Provided theoretical foundation for LR applications
Chen *et al.* (2023)[8]	Disease prediction with big data	ML over healthcare data	Applied ML techniques for large-scale disease prediction

Abbreviations: AI: Artificial intelligence; LR: Logistic regression; ML: Machine learning; SHAP: Shapley Additive Explanations; XGBoost: Extreme gradient boosting.

the final prediction depends on the average probability projection across models. Such an approach is especially applicable in disease risk assessment, where multiple models can be used to estimate a patient's probability of developing a particular disease, and an ensemble combines those scores to produce a final risk prediction. Ensemble learning uses an aggregation mechanism to combine the predictions of multiple models. Majority voting is usually employed in classifiers, where the output is the prediction of the majority of models, with several variations. Simple or weighted averaging is commonly used to compute regression predictions. Weighted averaging can give more weight to trustworthy models in the final decision, based on performance analysis. There are several advantages of ensemble learning over single-model methods. One of the primary advantages of ensemble learning is its ability to improve predictive accuracy by combining complementary information from multiple models, as collaborating models can counterbalance the biases and errors of any individual model. Second, it improves generalization: that is, it reduces the chance that the model overfits the training data and can handle unseen data. Third, ensemble methods reduce variability, enabling more stable predictions even on noisy or incomplete data. In medical use, datasets can be imbalanced, missing values are common, and patients with the same condition can exhibit different symptoms. Furthermore, ensemble learning can handle complex, high-dimensional healthcare datasets more effectively than many individual classifiers, including EHR, genetic, and multimodal health datasets. Ensemble models have gained popularity in the prediction of chronic diseases, including heart disease, diabetes, CKD, Alzheimer's disease, and cancer.[35] Learning the multitude of system structures for diagnosing chronic diseases typically follows a structured pipeline comprising data preprocessing, feature selection, model selection, ensemble methods, and, lastly, prediction generation. Data preprocessing is the initial step of the process: data cleaning and removal of redundancy and noise from raw medical data acquired from EHRs, imputation of missing data, and normalization or standardization of data to enhance model efficiency.[36]

The feature selection process aims to remove irrelevant or redundant features, thereby reducing computational effort to a large extent while maintaining predictive performance. RFE, information gain, Chi-square test, principal component analysis, and local RFE are typical feature selection methods. Selecting the most pertinent features enables ensemble models to work with relevant

attributes, thereby improving diagnostic accuracy. After feature selection, base learners are trained using multiple machine learning models. Examples of these models include random forests, XGBoost, SVMs, k-nearest neighbors (KNNs), artificial neural networks, and logistic regression. Once the out-of-the-box models are trained, a suitable ensemble technique like bagging, boosting, stacking, or voting is used to combine the predictions. Lastly, the ensemble generates diagnostic output by predicting whether a patient is vulnerable to a given chronic illness.[37] Analysis has shown that ensemble learning significantly improves the prediction of chronic diseases. For example, experiments have demonstrated that stacking models composed of random forests, XGBoost, and neural networks achieve greater precision in diagnosing CKD than the separate models. On the same note, the use of modified boosting algorithms such as XGBoost and LightGBM for predicting diabetes and heart disease has been reported to be more effective than conventional classifiers, as they aim to capture feature interactions and the complex relationships among risk factors.[38] One study found that a combination of ensemble models, including logistic regression, SVMs, and decision trees, outperformed individual classifiers in predicting heart failure. Moreover, XAI methods (e.g., SHAP, LIME) are being incorporated into ensemble structures to make models more transparent. That is essential in healthcare, where practitioners need to understand the context in which a model makes a particular prediction. Ensemble models are commonly optimized using Bayesian optimization and hyperparameter tuning techniques, such as grid search, random search, and genetic algorithms. Ensemble learning has its drawbacks, although it is beneficial.[39] Computational complexity is a major challenge, as training numerous models consumes more resources than training a single model. Ensemble methods can be hard to interpret, particularly when employing sophisticated algorithms such as XGBoost or neural networks. That is why recent progress in XAI has focused on interpreting ensemble models. Another problem is overfitting, particularly in constrained datasets; combining many models can lead to excessively complex choice boundaries.[40] This problem is commonly alleviated with methods such as cross-validation, regularization, and trimming. The outlook for ensemble learning in the diagnostics of chronic illnesses is promising, with deep learning, federated learning, and reinforcement learning advancing to increase predictive capabilities. It is expected that combining multimodal data, such as genetic information, medical imaging, wearable device data, and clinical reports, with ensemble models can lead to earlier diagnosis and more personalized therapy recommendations. Ensemble learning systems operational in hospitals and health institutions in real time would enable the real-time monitoring of patients, leading to timely interventions and the management of illnesses.[41] A learning ensemble framework is an exceptionally efficient and stable approach to chronic illness identification, leveraging the potential of multiple models to further enhance precision, mitigate overdependence, and improve decision-making. Ensemble techniques offer significant advantages over single-model techniques, as they incorporate methods such as bagging, boosting, stacking, and voting alongside feature selection, hyperparameter optimization, and explainability.[42] As machine learning, AI, and big data analytics continue to advance, ensemble frameworks are likely to become even more significant for achieving improved healthcare outcomes and facilitating accurate, early, and reliable diagnosis of chronic diseases[43] (Figure 1).

The above framework ensures a systematic approach to data preprocessing, feature selection, model training, and evaluation, making it suitable for predictive healthcare applications.

Table 5. Major heart disease risk factors

Risk factor	Clinical significance
Age	Risk increases with age
Chest pain type	Strong indicator of cardiac abnormalities
Resting blood pressure	Associated with hypertension
Cholesterol level	Linked to coronary artery disease
Maximum heart rate	Reflects cardiovascular fitness
Exercise-induced angina	Indicates myocardial ischemia
St depression	Marker of abnormal ECG response
Coronary arteries (CA)	Indicates vessel blockage severity
Thalassemia	Associated with cardiovascular complications

Abbreviation: ECG: Electrocardiography.

Figure 1. The proposed framework

Abbreviations: ECG: Electrocardiography; ROC–AUC: Receiver operating characteristic area under the curve; SHAP: Shapley Additive Explanations.

5. Proposed algorithms

Heatmaps are a mandatory visualization method for examining correlations and trends in EHR data. They help identify relationships among the various clinical variables that are essential to the initial evaluation of chronic illnesses. The suggested algorithm runs well on EHR data, calculates feature correlations, and produces a heatmap that is easy to interpret visually. As a first step in the algorithm, data preprocessing should be performed: It involves removing/encoding non-numeric columns (e.g., patient IDs or categorical variables). A looping system is then programmed to iterate over all numerical features, calculating pairwise correlations and producing a correlation matrix. It is then incorporated into the heatmap using visualization libraries such as Seaborn and Matplotlib, ensuring a clear presentation of variable interactions. To improve readability, the algorithm uses personalized loops for the axes' labels, color scheme, and annotations. The suggested algorithm has a significant advantage, especially for large-scale EHR data, where looping concepts allow processing without missing critical feature dependencies needed to enhance predictive modeling of chronic disease diagnosis by healthcare professionals (Table 6).

Table 6. Proposed algorithms

Algorithms	Details

Algorithm 1: Data preprocessing	Begin
	Load EHR dataset
	For each column in dataset:
	  If missing values exist:
	  If column is numerical:
	  Fill missing values with mean
	  Else if column is categorical:
	  Fill missing values with mode
	For each categorical feature in dataset:
	  Apply One-Hot Encoding
	For each numerical feature in dataset:
	  Normalize values using Min-Max scaling
	Split dataset into training (80%) and testing (20%) sets
	End

Algorithm 2: Training multiple base models	Begin
	Initialize base models: Decision Tree, Random Forest, Gradient Boosting
	For each model in base models:
	  Train model using training dataset
	  Store trained model
	End

Algorithm 3: Ensemble prediction using averaging (regression)	Begin
	Input: Trained models, test dataset
	Initialize result list
	For each test instance in test dataset:
	  Initialize sum_predictions = 0
	  For each model in trained models:
	  Predict numerical value
	  Add prediction to sum_predictions
	  Compute final prediction as sum_predictions / total_models
	  Append final prediction to result list
	Return result list
	End

Algorithm 4: Model evaluation	Begin
	Input: True labels, Predicted labels
	Compute accuracy, precision, recall, and F1-score
	If accuracy < threshold:
	  For each model in trained models:
	  Tune hyperparameters and retrain model
	Return evaluation metrics
	End

Algorithm 5: Generate heatmap for feature correlation in EHR data	Begin
	Load EHR dataset
	Remove non-numeric columns (if necessary)
	
	Initialize correlation_matrix
	For each feature X in dataset:
	  For each feature Y in dataset:
	  Compute correlation between X and Y
	  Store value in correlation_matrix
	
	Import heatmap visualization library (e.g., Seaborn, Matplotlib)
	Plot heatmap using correlation_matrix
	Customize color map and labels for better readability
	Display heatmap
	End

Algorithm 6: Predict disease risk on new EHR data	Begin
	Input: New patient EHR data
	Preprocess new data (apply same steps as training data)
	Initialize predictions list
	For each trained model:
	  Predict disease risk
	  Append prediction to predictions list
	Compute final prediction using majority voting (classification) or averaging (regression)
	Return final disease prediction with probability score
	End

Abbreviation: EHR: Electronic Health Records.

6. Experimental setup

The Python software system used packages such as Scikit-learn, Pandas, NumPy, Seaborn, and Matplotlib for data processing, model training, and visualization. Various machine learning algorithms, including decision trees, random forests, gradient boosting, SVMs, and neural networks, were developed and tested to ensure robustness. A comparative study was conducted only on a selected set of models based on their efficiency, accuracy, and comprehensibility. Our proposed ensemble model was then compared with these benchmark models, demonstrating its effectiveness in identifying chronic diseases in a timely fashion. The results supported the claim that our approach promoted accurate prediction and provided reliable information on illness trends.

7. Results

7.1. Data description

The study used the publicly available University of California, Irvine (UCI) heart disease dataset, which contains 303 patient records, 13 clinical attributes, and a binary target variable indicating the presence or absence of heart disease. This dataset is important for early detection of chronic illnesses, as it contains essential clinical and demographic data that influence patients' health. The data have numerous features; to select features and create the model, we trained it on a subset of the most pertinent ones. This dataset consists of cardiovascular markers essential for diagnosing conditions such as coronary artery disease, hypertension, and heart failure.

The key features used in our study were:

- Age: Represents the patient's age, a critical factor in assessing the risk of chronic diseases.
- Resting blood pressure: Measures resting blood pressure in mmHg, helping identify hypertension and cardiovascular risk.
- Cholesterol: Represents the patient's cholesterol levels, which are crucial for detecting heart disease risk.
- Fasting blood sugar: Indicates whether the patient's fasting blood sugar level exceeds a threshold, a key metric for diabetes risk assessment.
- Max heart rate: Represents the maximum recorded heart rate during a stress test, which helps assess cardiovascular efficiency and overall heart health.
- Exercise-induced angina: A binary feature indicating whether the patient experienced chest pain due to exercise, an important indicator of heart disease.
- ST depression: Measures ST-segment depression in electrocardiography (ECG) readings, providing insight into abnormal heart function and ischemia risks.

The PROMISE dataset, an open-source EHR dataset containing important clinical factors for predicting chronic diseases, serves as the basis for this paper, as shown in Figure 2. Data visualization of this dataset comprises numerous statistical plots that facilitate interaction and interpretation of feature distributions, correlations, and patterns critical to feature selection.

The age histogram reveals that the majority of the patients fall under the 50–60 age bracket, and therefore chronic diseases are a major risk factor. The sex distribution indicates that the dataset contains more male than female patients, suggesting a gender imbalance. The histogram of the chest pain type shows various types of angina, and the most frequent distress types were 0 and 2.

These are very common and important for identifying risks associated with heart disease. The histogram of resting blood pressure shows a mild right tail, with the majority of patients having plasma pressure of 120–140 mmHg, and some with very high values, a factor highly indicative of hypertension. The scatter plots also tested relationships among important variables, such as age vs. sex, which showed they are independent, and sex vs. chest pain type, which showed differences in the number of chest pain episodes between men and women. The scatter plot of chest pain type vs. resting blood pressure shows that certain chest pain types are associated with elevated blood pressure, strengthening their significance in disease diagnosis.

This shows that our ensemble learning approach, feature engineering, and hyperparameter optimization are superior to the other systems; they are more robust and efficient at predicting heart disease.

Figure 3 presents an exploratory data analysis (EDA) for heart disease prediction, displaying the distribution of key categorical features. The chest pain type distribution shows that most patients had Type 0 (asymptomatic), while Types 1–3 occurred less frequently, which is crucial since certain chest pain types indicate higher cardiovascular risk. The resting ECG distribution shows that most patients had normal (0) or slightly abnormal (1) ECGs, with very few in the severe abnormality category (2). The ST slope distribution shows that most patients have an upsloping or flat ST segment, whereas a downsloping ST segment (0), a strong indicator of ischemia, is rare. The coronary artery distribution shows that most patients had no major coronary blockages (0 colored vessels), but as the number of blocked arteries increased (1–4), the risk of severe heart disease also increased. The thalassemia distribution suggests that most patients fall into categories 2 and 3, with fewer in category 1 and very few in category 0, indicating varying levels of blood disorders affecting

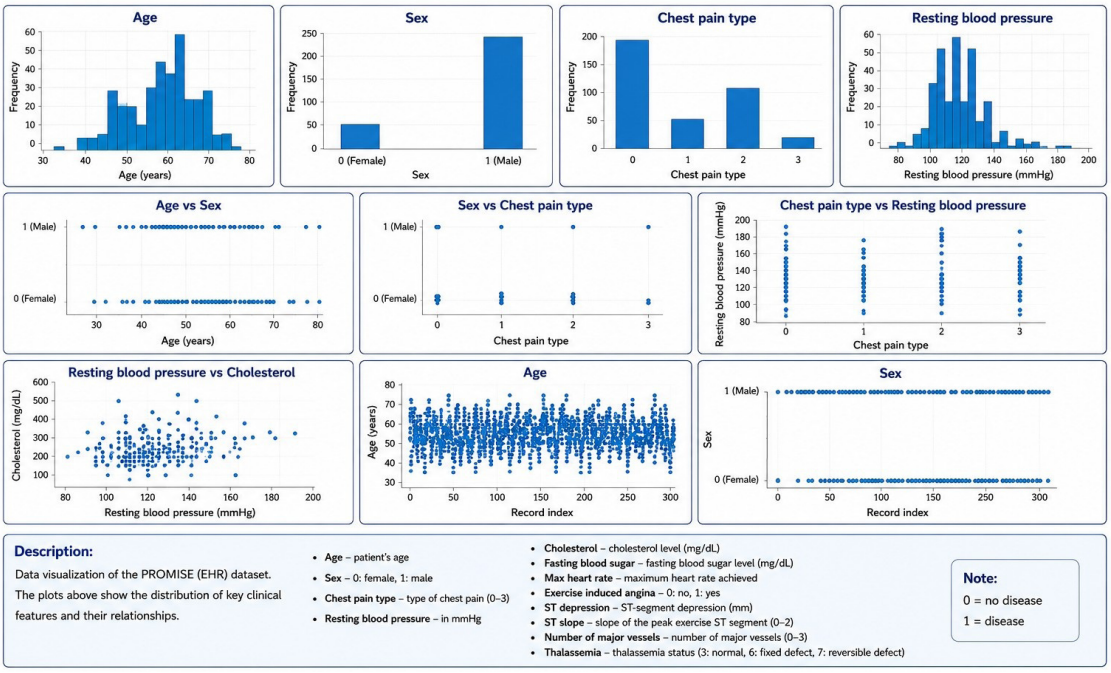

Figure 2. Feature visualization. (A) Age. (B) Sex. (C) Chest pain type. (D) Resting blood pressure. (E) Age vs. sex. (F) Sex vs. chest pain type. (G) Chest pain type vs. resting blood pressure. (H) Resting blood pressure vs. cholesterol. (I) Age by record index. (J) Sex by record index.

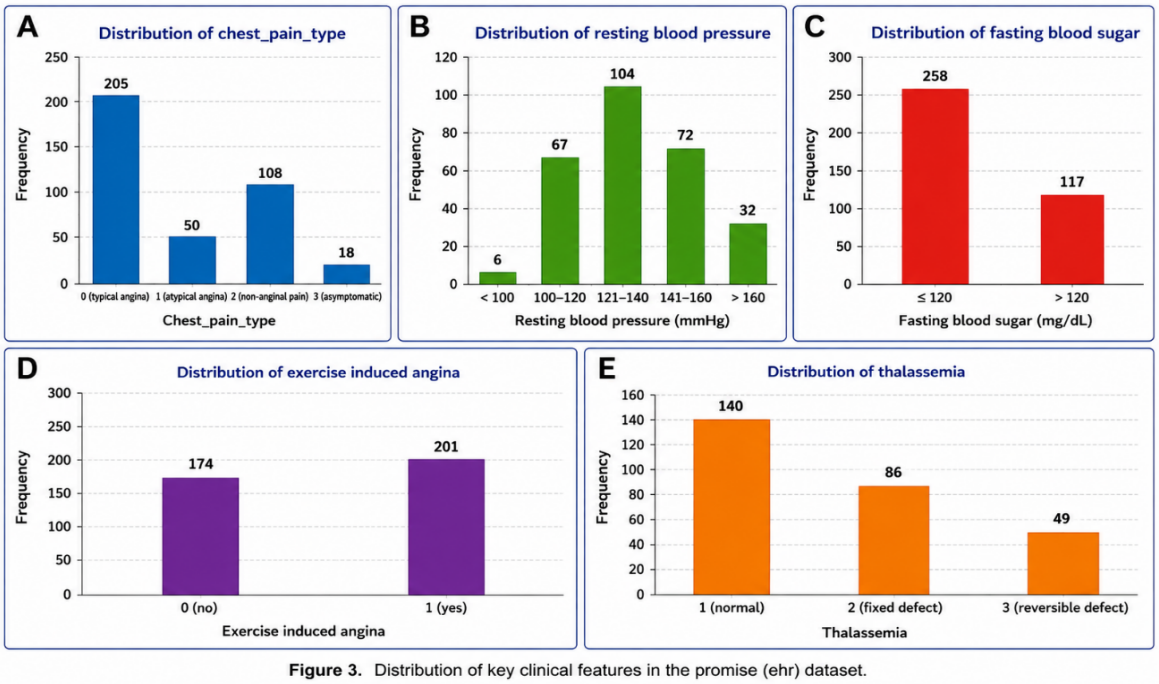

Figure 3. Exploratory data analysis for heart disease prediction. (A) Distribution of chest pain type. (B) Distribution of resting electrocardiography (ECG). (C) Distribution of ST slope. (D) Distribution of coronary arteries. (E) Distribution of thalassemia.

oxygen transport. These insights highlight potential class imbalances in the dataset and confirm that chest pain type, ST slope, and coronary artery blockages are highly predictive of heart disease. This EDA helps refine feature selection, ensuring that only the most relevant clinical attributes are used in the machine learning model for early heart disease detection.

The correlation heatmap (Figure 4) provides a detailed analysis of the relationships between various clinical features in the dataset and their influence on heart disease prediction. The strongest positive correlation with heart disease was observed in chest pain type (0.43), indicating that certain types of chest pain significantly increase the risk. Similarly, exercise-induced angina (−0.44) and coronary arteries (−0.38), which represent angina during exertion and the number of major coronary blockages detected via fluoroscopy, respectively, also showed strong correlations with heart disease. Conversely, the ST slope (0.35) shows the strongest negative correlation, indicating that a downsloping ST segment in ECG readings is a strong predictor of heart disease. Other notable correlations included maximum heart rate (0.42), indicating that patients with lower maximum heart rates during exercise are more prone to heart disease, and age (−0.23), suggesting that older individuals generally have a higher risk. Interestingly, resting blood pressure (−0.14) and cholesterol (−0.10) exhibited weak correlations, implying that these commonly measured cardiovascular indicators alone may not be strong predictors of heart disease. Sex (−0.28) had a weak negative correlation, suggesting that gender alone is not a major determinant of heart disease risk. The heatmap also highlights multicollinearity between certain features, such as ST depression vs. ST slope (−0.58) and age vs. maximum heart rate (−0.39), meaning that including both in a predictive model may introduce redundancy. Based on these insights, chest pain type, ST slope, coronary arteries, exercise-induced angina, maximum heart rate, and age emerge as the most significant features for heart disease prediction, while features with weak correlations may be less impactful. This correlation heatmap helps refine feature selection, ensuring that only the most relevant clinical variables are used in machine learning models, thereby improving accuracy and efficiency in early heart disease detection (Figure 5).

Figure 6 presents the comparison of the classification accuracies achieved by various machine learning models with all available features and with mutual information-based feature selection. The results demonstrate that feature selection consistently improved model performance across all evaluated classifiers. By removing redundant and less informative attributes, the feature-selection process enabled the models to focus on the most relevant clinical variables associated with heart disease diagnosis.

Among the baseline classifiers, the accuracy of logistic regression improved from 84.39% to 86.21%, while that of

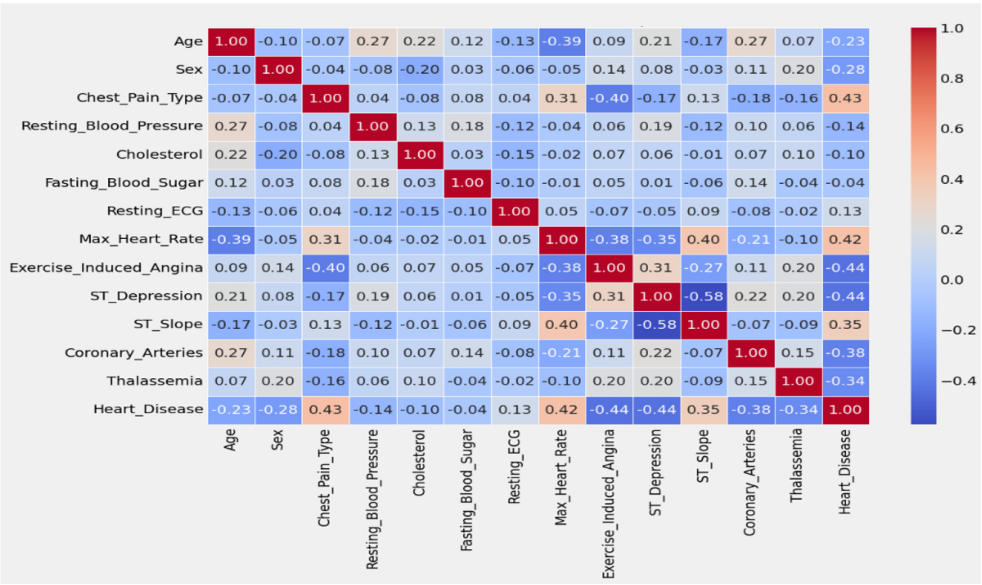

Figure 4. Correlation heat map
Abbreviation: ECG: Electrocardiography.

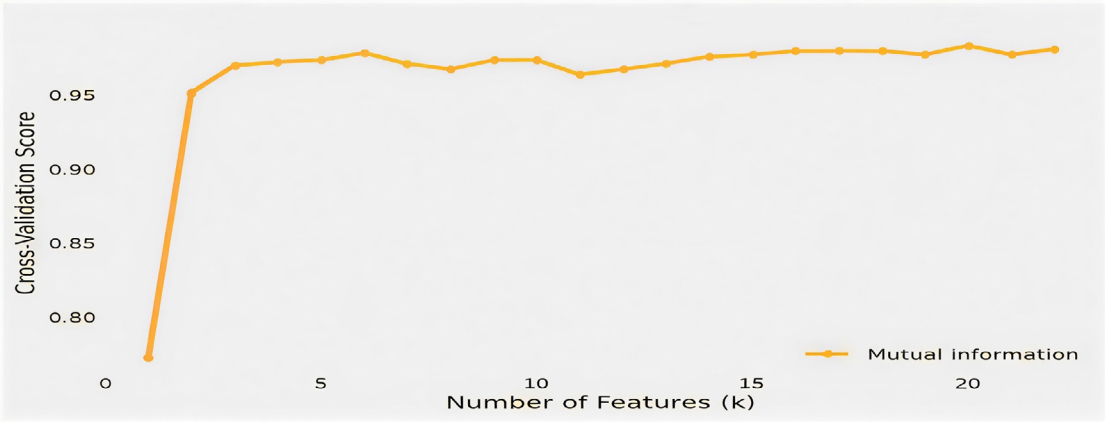

Figure 5. Selection of the best number of features in mutual information

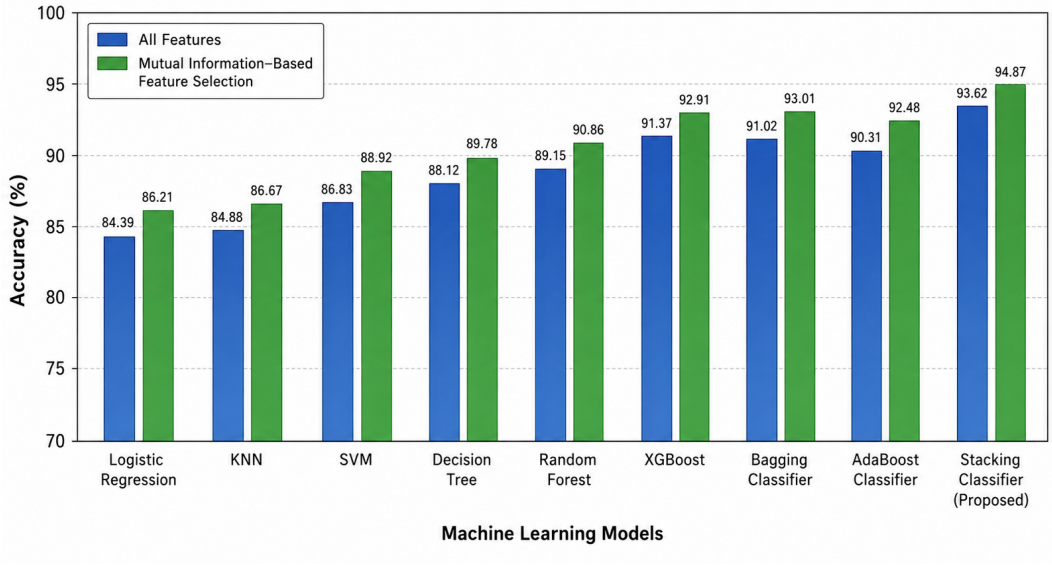

Figure 6. Comparison of classification accuracy obtained using all features and SVM–RFE-based feature selection across different machine learning models

KNN increased from 84.88% to 86.67%. Similarly, SVM accuracy improved from 86.83% to 88.92%, indicating that the selected features enhanced class discrimination. Tree-based models also benefited from feature selection, with decision tree accuracy increasing from 88.12% to 89.78% and random forest improving from 89.15% to 90.86%.

The most significant performance gains were observed in ensemble learning models. XGBoost achieved an accuracy of 92.91% after feature selection, compared to 91.37% with all features. Bagging and AdaBoost classifiers also demonstrated improved performance, achieving accuracies of 93.01% and 92.48%, respectively. The proposed stacking classifier achieved the highest overall performance, increasing accuracy from 93.62% to 94.87% and outperforming all other models.

These findings confirm that feature selection plays a crucial role in improving classification accuracy by reducing feature redundancy, minimizing noise, and enhancing

model generalization. The superior performance of the proposed stacking-based ensemble framework further demonstrates the effectiveness of combining multiple classifiers with optimized feature subsets for heart disease diagnosis.

Table 7 presents the assessment metrics of machine learning models incorporating all characteristics in the proposed algorithmic heart disease prediction strategy. Because it captured complex decision boundaries, the proposed ensemble model achieved 93.62% accuracy, 94.08% precision, 93.11% recall, and 93.58% F1-score, demonstrating its classification capacity. SVM achieved 86.83% accuracy, with similarly high precision (87.42%) and recall (86.83%), indicating strong predictive performance in distinguishing heart disease cases. KNN and logistic regression also performed well, with accuracies of 84.88% and 84.39%, respectively, and similar precision, recall, and F1 scores, showcasing their reliability in classification. These results validate that the proposed ensemble learning and preprocessing steps significantly enhance model performance, ensuring high accuracy and effective disease prediction. The high recall values across all models indicate that the approach minimizes false negatives, a crucial aspect of medical diagnosis. Overall, the proposed methodology proves to be highly effective in building a reliable heart disease prediction system.

7.2. Receiver operating characteristic curve analysis

Figure 7 illustrates the ROC curves of the evaluated classifiers. The ROC curve summarizes the trade-off between sensitivity and specificity across classification thresholds. A curve closer to the top-left corner indicates stronger discriminative power. In this study, the best-performing model showed the largest ROC–AUC, confirming superior separation between patients with and without heart disease. The ROC analysis supports the findings in Table 7 and demonstrates that the proposed framework is effective not only at a fixed threshold but also across different decision thresholds.

7.3. Confusion matrix of the best model

Figure 8 shows the confusion matrix of the best-performing model. The matrix indicates the number of true positives, true negatives, false positives, and false negatives. The high number of correct predictions and the very low number of misclassifications confirm that the model is highly effective for heart disease diagnosis. In a clinical setting, minimizing false negatives is especially important because missed cases may delay treatment. The confusion-matrix results therefore suggest that the model has strong potential as a decision-support tool for early screening.

7.4. k-fold cross-validation results

Table 8 reports the results of k-fold cross-validation used to assess the stability of the proposed framework. Cross-validation provides a more reliable estimate of generalization performance than a single train-test split because it evaluates the model across multiple data partitions. The results indicate that model performance remains stable across folds, suggesting that the proposed approach is not overly dependent on a specific dataset split.

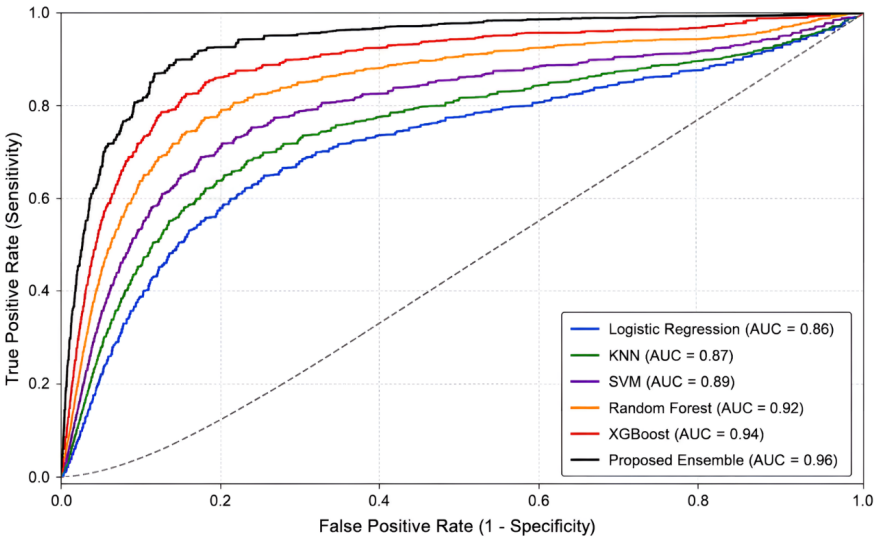

Figure 7. Receiver operating characteristic curves for the evaluated models
Abbreviations: AUC: Area under the curve; KNN: k-nearest neighbor; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

Table 7. Performance comparison of machine learning models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
KNN	84.88	85.28	84.88	84.83	0.87
Logistic regression	84.39	85.47	84.39	84.26	0.86
SVM	86.83	87.42	86.83	86.77	0.89
Random forest	89.15	89.84	88.73	89.21	0.92
XGBoost	91.37	91.85	91.02	91.41	0.94
Proposed ensemble model	93.62	94.08	93.11	93.58	0.96

Abbreviations: KNN: k-nearest neighbor; ROC–AUC: Receiver operating characteristic area under the curve; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

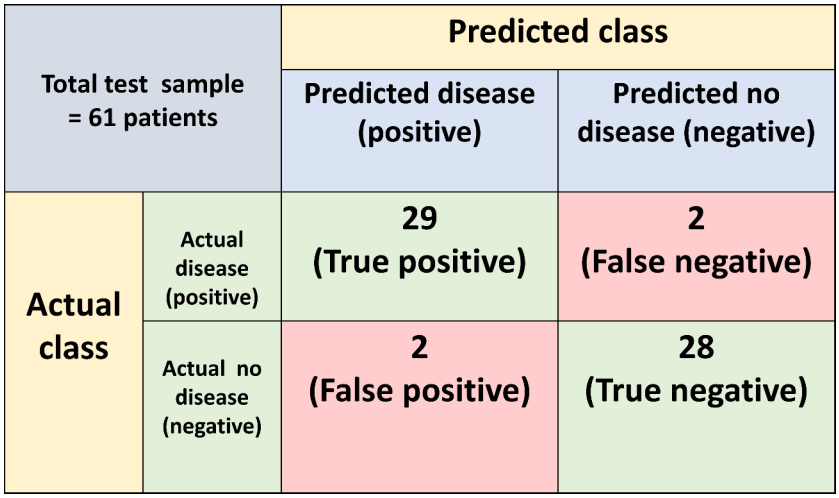

Figure 8. Confusion matrix of the best-performing model

Table 8. Stratified five-fold cross-validation results of the proposed ensemble model

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Fold 1	93.12	93.45	92.96	93.20
Fold 2	94.26	94.71	93.85	94.28
Fold 3	92.87	93.12	92.65	92.88
Fold 4	94.01	94.35	93.74	94.04
Fold 5	93.85	94.20	93.40	93.80
Mean ± SD	93.62 ± 0.58	93.97 ± 0.64	93.32 ± 0.51	93.64 ± 0.57

8. Conclusion

This study proposed a hybrid feature-selection and ensemble learning framework for early heart disease diagnosis using EHR data. The framework combined SVM–RFE for feature selection, Bayesian optimization for hyperparameter tuning, and multiple machine learning classifiers to improve predictive performance and model robustness. The experimental results showed that the proposed ensemble model achieved 93.62% accuracy, 94.08% precision, 93.11% recall, 93.58% F1-score, and 0.96 ROC–AUC, confirming that the proposed approach outperforms the baseline classifiers used in this study. The selected features, including chest pain type, ST slope, exercise-induced angina, maximum heart rate, and age, were found to be clinically relevant and highly informative for heart disease prediction. The cross-validation results further demonstrated stable performance and good generalization. The proposed framework can therefore serve as an interpretable decision-support tool for early screening and risk assessment in clinical practice. Future work should focus on external validation using multi-center datasets, advanced XAI techniques, and integration with real-time clinical decision-support systems.

The proposed framework demonstrates strong potential for clinical decision support by identifying patients at elevated cardiovascular risk using routinely collected EHR data. The incorporation of SVM–RFE-based feature selection improves model interpretability by highlighting clinically meaningful predictors, while Bayesian Optimization and ensemble learning enhance robustness and generalization. These characteristics make the framework suitable for supporting early screening and risk assessment in real-world healthcare environments.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors have no competing interests to declare that are relevant to the content of this article.

Authors Contribution

Conceptualization: All authors
Formal analysis: All authors
Methodology: Gaurav Dwivedi, Anurag Tripathi, Rahul Vyas
Visualization: Gaurav Dwivedi, Anurag Tripathi, Rahul Vyas
Writing–original draft: Gaurav Dwivedi, Anurag Tripathi, Rahul Vyas
Writing–review & editing: Gaurav Dwivedi, Anurag Tripathi, Rahul Vyas

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The datasets used and/or analyzed during the current study are publicly available and can be provided by the corresponding author upon reasonable request.

References

1. Dwivedi G, Vyas R, Tripathi A, *et al.* MRI-Driven Deep Learning in Parkinson's Disease Diagnosis: Innovations, Challenges, and Future Directions. *Proc Natl Acad Sci India Sect A Phys Sci*. 2026.

   doi: 10.1007/s40010-026-01003-7

2. Manjulatha B, Pabboju S. An ensemble model for predicting chronic diseases using machine learning algorithms. In: Satapathy SC, Bhateja V, Favorskaya MN, Adilakshmi T, eds. *Smart Computing Techniques and Applications. Smart Innovation, Systems and Technologies*. Singapore: Springer; 2021: vol 224.

   doi: 10.1007/978-981-16-1502-3_34

3. Al-Jamimi HA. Synergistic feature engineering and ensemble learning for early chronic disease prediction. *IEEE Access*. 2024;12:62215-62233.

   doi: 10.1109/access.2024.3395512

4. Burse K, Kirar VPS, Burse A, Burse R. Various preprocessing methods for neural network-based heart disease prediction. In: Tiwari S, Trivedi M, Mishra K, Misra A, Kumar K, eds. *Smart Innovations in Communication and Computational Sciences. Advances in Intelligent Systems and Computing*. Singapore: Springer; 2019: vol 851.

   doi: 10.1007/978-981-13-2414-7_6

5. Zhou H, Zhang PY, Zou X, Liu J, Wang WJ. Chronic disease diagnosis model based on convolutional neural network and ensemble learning method. *Digital Health*. 2023;9:20552076231198643.

   doi: 10.1177/20552076231198643

6. Listiana E, Muzayanah R, Muslim MA, Sugiharti E. Optimization of support vector machine using information

gain and AdaBoost to improve accuracy of chronic kidney disease diagnosis. *J Soft Comput Explor.* 2023;4(3):152-158.

doi: 10.52465/joscex.v4i3.218

7. Al-Jamimi HA, BinMakhashen GM, Saleh TA. Artificial intelligence approach for modeling petroleum refinery catalytic desulfurization process. *Neural Comput Appl.* 2022;34(20):17809-17820.

doi: 10.1007/s00521-022-07423-x

8. Chen H. Enterprise marketing strategy using big data mining technology combined with XGBoost model in the new economic era. *PLoS ONE*. 2023;18(6):e0285506.

doi: 10.1371/journal.pone.0285506

9. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. Wiley Series in Probability and Statistics. New York, NY, USA: John Wiley & Sons, Inc; 2013.

doi: 10.1002/9781118548387

10. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res.* 2012;13(1):281-305.

11. Wu J, Chen XY, Zhang H, Xiong LD, Lei H, Deng S. Hyperparameter optimization for machine learning models based on Bayesian optimization. *J Electron Sci Technol.* 2019;17(1):26-40.

doi: 10.11989/JEST.1674-862X.80904120

12. Jongbo OA, Adetunmbi AO, Ogunrinde RB, Badeji-Ajisafe B. Development of an ensemble approach to chronic kidney disease diagnosis. *Sci Afr.* 2020;8:e00456.

doi: 10.1016/j.sciaf.2020.e00456

13. Hannan SA, Pal P. Detection and classification of kidney disease using convolutional neural networks. *J Neurol Neurorehab Res*. 2023;8(2):136.

doi: 10.35841/aajnnr-8.2.136

14. Chicco D, Lovejoy CA, Oneto L. A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease. *IEEE Access*. 2021;9:165132-165144.

doi: 10.1109/access.2021.3133700

15. Moreno-Sánchez PA. Data-Driven Early Diagnosis of Chronic Kidney Disease: Development and evaluation of an explainable AI model. *IEEE Access*. 2023;11:38359-38369.

doi: 10.1109/access.2023.3264270

16. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2016:785-794.

doi: 10.1145/2939672.2939785

17. Betrò B. Bayesian Methods in Global Optimization. In: Gritzmann P, Hettich R, Horst R, Sachs E, eds. *Operations Research '91*. Heidelberg, Germany: Physica-Verlag HD. 1992:16-18.

doi: 10.1007/978-3-642-48417-9_6

18. Jain V, Agrawal M. Heart Failure Prediction Using XGB Classifier, Logistic Regression and Support Vector Classifier. In: Proceedings of the 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT). New York, NY, USA: IEEE; 2023:1-5.

doi: 10.1109/incacct57535.2023.10141752

19. Kong Y, Wang Y, Sun S, Wang J. XGB and SHAP credit scoring model based on Bayesian optimization. *J Comput Electron Inf Manag.* 2023;10(1):46-53.

doi: 10.54097/jceim.v10i1.5368

20. Basha N, Ibrahim G, Choudhury HA, *et al.* Bayesian-optimized Neural Networks and their application to model gas-to-liquid plants. *Gas Sci Eng.* 2023;113:204964.

doi: 10.1016/j.jgsce.2023.204964

21. Sharmila SL, Dharuman C, Venkatesan P. Disease classification using machine learning algorithms-a comparative study. *Int J Pure Appl Math.* 2017;114(6):1-10.

22. Ruiz VM, Goldsmith MP, Shi L, *et al.* Early prediction of clinical deterioration using data-driven machine-learning modeling of electronic health records. *J Thorac Cardiovasc Surg.* 2022;164(1):211-222.e3.

doi: 10.1016/j.jtcvs.2021.10.060

23. Liu T, Krentz A, Lu L, Wang Y, Curcin V. Benchmarking survival machine learning models for 10-year cardiovascular disease risk prediction using large-scale electronic health records. *Digital Health*. 2026;12.

doi: 10.1177/20552076251408534

24. Zaman MAU, Adepoju OG. A Solution Approach with Ensemble-Based Learning Technology for Predicting Early Readmission Among Patients with Heart Failure (HF) Diagnosis Using Electronic Health Records (EHR). *Medinformatics*. 2025.

doi: 10.47852/bonviewmedin52025119

25. S NV, T M, Devi SP. Integrating Machine Learning and Deep Learning Approaches for Accurate Cardiovascular Disease Prediction from Electronic Health Records. In: Proceedings of the 2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI). India; Erode; 2025:1090-1096.

doi: 10.1109/icmsci62561.2025.10894568

26. Sowmiya R, Devi BR, Vasuki R, *et al.* Combining Deep Learning and Machine Learning Techniques to Predict Cardiovascular Disease Accurately from Electronic Health Records. In: Proceedings of the 2025 3rd International Conference on Sustainable Computing and Data

Communication Systems (ICSCDS). New York, NY, USA: IEEE; 2025:466-470.

doi: 10.1109/icscds65426.2025.11166930

27. Pathan MS, Nag A, Pathan MM, Dev S. Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthc Anal.* 2022;2:100060.

doi: 10.1016/j.health.2022.100060

28. Sarra RR, Dinar AM, Mohammed MA, Abdulkareem KH. Enhanced Heart Disease Prediction Based on Machine Learning and χ2 Statistical Optimal Feature Selection Model. *Designs*. 2022;6(5):87.

doi: 10.3390/designs6050087

29. Noroozi Z, Orooji A, Erfannia L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Sci Rep*. 2023;13(1).

doi: 10.1038/s41598-023-49962-w

30. Dissanayake K, Md Johar MG. Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms. *Appl Comput Intell Soft Comput.* 2021;2021:1-17.

doi: 10.1155/2021/5581806

31. Manikandan G, Pragadeesh B, Manojkumar V, Karthikeyan AL, Manikandan R, Gandomi AH. Classification models combined with Boruta feature selection for heart disease prediction. *Inform Med Unlocked.* 2024;44:101442.

doi: 10.1016/j.imu.2023.101442

32. Ara J, Bhuiyan H, Roza II, Nahin ASM. Importance of balanced datasets with feature selection and ensemble methods on heart disease classification using distinctive machine learning techniques: a comparative analysis. *Sci Rep*. 2026;16(1):11706.

doi: 10.1038/s41598-026-47691-4

33. Luo L, Wang Y, Mo DY. Identifying heart disease risk factors from electronic health records using an ensemble of deep learning method. *IISE Trans Healthc Syst Eng.* 2023;13(3):237-247.

doi: 10.1080/24725579.2023.2205665

34. Dwivedi G, Narang T. A Comprehensive Review of Graph-Based Anomaly Detection Algorithms for Complex Data Environments. *RACSC*. 2025;19.

doi: 10.2174/0126662558383302250922190426

35. Bebortta S, Tripathy SS, Basheer S, Chowdhary CL. FedEHR: A Federated Learning Approach towards the Prediction of Heart Diseases in IoT-Based Electronic Health Records. *Diagnostics*. 2023;13(20):3166.

doi: 10.3390/diagnostics13203166

36. Malwade S, Sanap AR, Gite YS, Pandit P, Gaikwad YJ, Mhatre A. Predicting Heart Diseases in IoT-Based Electronic Health Records: A Federated Learning Approach. *J Electr Syst.* 2024;20(2):472-484.

doi: 10.52783/jes.1465

37. Gupta PK, Dalei V. Heart Disease Prediction Through Health Record Analysis Using Ensemble Learning. In: Proceedings of the 2025 OITS International Conference on Information Technology (OCIT). Bhubaneswar, India. 2025:1111-1116.

doi: 10.1109/ocit66168.2025.11400391

38. Jrab D, Eleyan D, Eleyan A, Bejaoui T. Heart Disease Prediction Using Machine Learning Algorithms. In: Proceedings of the 2024 International Conference on Smart Applications, Communications and Networking (SmartNets). VA, USA: Harrisonburg; 2024:1-8.

doi: 10.1109/smartnets61466.2024.10577725

39. Sondos Jameel Mukhyber. Classification Of Heart Disease Using Feature Selection and Machine Learning Techniques. *Phys Sci Life Sci Eng.* 2025;2(3):9.

doi: 10.47134/pslse.v2i3.386

40. El-Sofany HF. Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques. *IEEE Access*. 2024;12:106146-106160.

doi: 10.1109/access.2024.3437181

41. Stevens C, Mahani A, Ray K, Vallejo-Vaz A, Sharabiani M. Ensemble machine learning for screening cardiovascular diseases in electronic health records. *Atherosclerosis*. 2023;379:S194.

doi: 10.1016/j.atherosclerosis.2023.06.644

42. Viswateja K, Supriya D, Jahnavi KB, Aruna VBKL. Enhanced Feature Selection Technique Using Machine Learning Algorithm for Accurate Heart Disease Classification. In: Proceedings of the 2025 International Conference on Computing Technologies (ICOCT). Bengaluru, India. 2025:1-6.

doi: 10.1109/icoct64433.2025.11118525

43. Aşkın L, Polat E, Hoşoğlu Y, Tanrıverdi O. The Application of Artificial Intelligence in the Field of Cardiovascular Diseases Focuses on Both Diagnostic and Therapeutic Aspects. *Exp Appl Med Sci.* 2024;5(2):22-35.

doi: 10.46871/eams.1438927